CS674/INFO630: Advanced Language Technologies, Fall 2007, Lecture by Lillian Lee
Lecture 5 Guide: An Introduction to Probabilistic Retrieval

Alex Chao David Collins

September 11, 2007

1 Introduction

In the last lecture, we finished our discussion of pivoted document-length normalization. Recall that the reason behind that technique was that information retrieval (IR) systems using the vector space (VS) model with L_2 length normalization were shown to be biased toward retrieving short documents. Simply altering term frequency and inverse document frequency weights themselves (by coming up with the "correct" term weights based on individual terms alone, as opposed to collective term statistics) would not be sufficient to alleviate the bias, because you need information about the document as a whole in order to normalize the document vectors correctly.

Pivoted document-length normalization and the VS model paradigm as a whole are very empirically driven. In this lecture, we introduce the probabilistic retrieval paradigm of information retrieval, which is a much more theoretically inclined perspective. It has been developing in parallel to the VS paradigm, but, as we will see in later lectures, both seemed to have come to many of the same conclusions. We will be discussing the Robertson & Spärck Jones (RSJ) variant of probabilistic retrieval in this lecture. For a brief history of this topic, see C.J. van Rijsbergen's paper entitled *The emergence of probabilistic accounts of information retrieval* (2005). As we discuss the probabilistic paradigm, it is important that we compare and contrast with the VS model, both in spirit and in the techniques used. Although probabilistic IR comes primarily from principled and theoretical ideas, it uses both statistical estimation and empirical substitution at times.

2 Setup

Assume we have a set of m attribute random variables A_j that each correspond to a document characteristic. Unlike in the VS model, each of the attributes here will generalize to all sorts of ways to describe a document, not just the presence of terms (at least for now). Here are two examples:

- $A_1 = yes \equiv$ document contains both "car" and "Ithaca".
- $A_{17} = 14 \equiv$ document is 14 words long.

We are intentionally trying to be as general as possible, so that we avoid unnecessarily limiting our model in scope. It is important for the purpose of our presentation, however, that the range of each attribute (A_j) contains a distinguished value of 0 (or *no*), meaning the document does not exhibit this attribute. It will be clear why we need this later in the lecture.

We will represent a document d as $\vec{d}$, where $d[j]$ is the value of A_j for d. Note that (ignoring normalization issues) the VS model is a special case of this, where each of the attributes refers to a term. As far as scoring a document for "relevance", we will, again, intentionally keep the semantics slightly undetermined so that we do not unnecessarily lose generality.

Let R_q be a random variable with respect to query q. R_q refers to how well a document is relevant to the query q. For now, we will assume $R_q \in \{y, n\}$. This range, however, generalizes easily; it does not need to be binary.

We will thus rank the documents using the following probability:

$$\Pr(R_q = y \mid \vec{A} = \vec{d})$$

This choice requires some examination. Namely, why is this quantity described as a probability if a particular document can be classified as either relevant or not (probabilities of 1 or 0, respectively)? The probability is in fact due to, perhaps among other things, the following factors:

- The set of attributes is not uniquely specifying. That is, the attributes may effectively lead to classifications of the documents in which a single "bin" may contain both relevant and non-relevant documents.
- There may be variation among users who judge the relevance of a given document differently, or there may even be variation for a single user whose judgment changes with time or depends on a specific context.

It should be noted that Robertson & Spärck Jones do not use $\Pr(R_q = y \mid \vec{A} = \vec{d})$, but instead use the following ranking function:

$$\log\left(\frac{\Pr(R_q = y \mid \vec{A} = \vec{d})}{\Pr(R_q = n \mid \vec{A} = \vec{d})}\right)$$

A claim was made in lecture that both ranking functions produce the same result, with the former having more convenient mathematical properties. In any case, our overall claim is that this general probabilistic ranking method better matches the retrieval goal, though there are objections (see *Gordon & Lenk '92*).

3 Derivation

Assuming we rank using $\Pr(R_q = y \mid \vec{A} = \vec{d})$, there are still several challenges to consider, the first being that there are no relevance labels for the documents. Also, there is not much information regarding the particular attribute vectors.

We start by performing a Bayes' "flip" on the rank function in order to condition on the variable with fewer possible values (or "bins"):

$$\Pr(R_q = y \mid \vec{A} = \vec{d}) = \frac{Pr(\vec{A} = \vec{d} \mid R_q = y) \Pr(R_q = y)}{\Pr(\vec{A} = \vec{d})}$$

This flip appears to be worse than what we started with, since we have more unknowns, until we take into account *document independence*. A term of the equation is said to be *document independent* if it contributes equally to every document's score, thereby making it independent of any given document so that its omission still preserves the overall ranking of the documents. Such terms can be dropped from the equation without altering the relative ranks of the documents. In this case, the term $\Pr(R_q = y)$ is document independent. We are left with:

$$\frac{\Pr(\vec{A} = \vec{d} \mid R_q = y)}{\Pr(\vec{A} = \vec{d})} \tag{1}$$

At this point, we would like to break this function down further, as it still contains variables whose values are not observed (e.g. $R_q = y$). In particular, we try to assume a conditional independence between the elements of $\vec{A}$ and R_q simultaneously with an independence among the elements of $\vec{A}$. However, it can be argued that such an assumption implies logical inconsistencies (*Cooper '95*), so we turn instead to a type of linked independence assumption, which enables us to decompose the ranking function as follows:

$$k \prod_{j=1}^{m} \frac{\Pr(A_j = d[j] \mid R_q = y)}{\Pr(A_j = d[j])}, k > 0 \tag{2}$$

The constant value k is a factor that accounts for the skew from independence that the decomposition introduces. Given that we are concerned only about ranking our documents, the term falls out of the equation.

We continue trying to simplify the ranking function, as we are still hindered by the relevancy component. We proceed by factoring the quantity based on the appearance or absence of the terms in the query, since the query is the only "clue" we have regarding relevance:

$$\Rightarrow \prod_{j:q[j]\neq 0} \frac{\Pr(A_j = d[j] \mid R_q = y)}{\Pr(A_j = d[j])} \times \prod_{j:q[j]=0} \frac{\Pr(A_j = d[j] \mid R_q = y)}{\Pr(A_j = d[j])} \tag{3}$$

An examination of the second product provides a further reduction. Namely, this product concerns those attributes that are not observed in the query, and it is not clear that such an attribute is ever anti-correlated with relevance. In other words, attributes that do not appear in the query can be said to have the same distribution across relevant documents as over all documents. Under this assumption, we have the following equality:

$$\Pr(A_j = d[j] \mid R_q = y) = \Pr(A_j = d[j])$$

This assumption causes the second product to fall out of the ranking function, leaving us with:

$$\Rightarrow \prod_{j:q[j]\neq 0} \frac{\Pr(A_j = d[j] \mid R_q = y)}{\Pr(A_j = d[j])} \tag{4}$$

Again, we are unable to estimate the part of the equation containing $R_q = y$, so we attempt a further reduction, this time factoring according to the attributes exhibited by the document. We arrive at the following ranking function:

$$\Rightarrow \prod_{j:q[j]\neq 0, d[j]\neq 0} \frac{\Pr(A_j = d[j] \mid R_q = y)}{\Pr(A_j = d[j])} \times \prod_{j:q[j]\neq 0, d[j]=0} \frac{\Pr(A_j = d[j] \mid R_q = y)}{\Pr(A_j = d[j])} \tag{5}$$

Let the first product be **I** and the second product, **II**. The second product is peculiar given that it concerns attributes that the document does not exhibit. If we could establish document independence for this product, it would drop out of the equation; at the moment, we can't, because the index of the product involves the document. We thus multiply the above ranking function by two additional product terms, **III** and **IV**, whose product is 1. **III** represents the missing terms from the ranking function so as to "fill out" the product index, and **IV** is simply its reciprocal.

$$\prod_{j:q[j]\neq 0, d[j]\neq 0} \frac{\Pr(A_j = 0 \mid R_q = y)}{\Pr(A_j = 0)}$$

$$\prod_{j:q[j]\neq 0, d[j]\neq 0} \frac{\Pr(A_j = 0)}{\Pr(A_j = 0 \mid R_q = y)}$$

The product **II** $\times$ **III** is a quantity that does not depend on the document, as none of the probabilities make any estimations using $\vec{d}$ and now the index of the combined product does not either. The ranking is thus unaffected by this product, and we are left with **I** $\times$ **IV**:

$$\prod_{j:q[j]\neq 0, d[j]\neq 0} \frac{\Pr(A_j = d[j] \mid R_q = y)}{\Pr(A_j = d[j])} \times \frac{\Pr(A_j = 0)}{\Pr(A_j = 0 \mid R_q = y)} \tag{6}$$

At this point, we have an arguably "nicer" ranking function that has been purged of document independent factors, but there still exists the relevancy component, which we have no easy way of estimating. The reader might now wonder if the sacrifices and assumptions made in the formulation of this ranking function were all worth it. The following lecture addresses this concern indirectly by taking the derivation even further.

4 Sample Questions

4.1 Question 1

We will now consider whether one of the assumptions we made about our attribute vector $\vec{A}$ limits the generality, and therefore the usefulness, of the ranking function.

4.1.1 Question 1 Part A

In the derivation, we needed to make the assumption that the range of each attribute A_j contains a distinguished value that means the attribute is "not on". Why was it necessary that we put such a restriction on the attribute vector?

4.1.2 Question 1 Part B

Does making this assumption mean that we cannot generalize our model to contain attributes that do not intuitively have a designated "not on" value such as $A_i \in \{$"document is handwritten", "document is typed"$\}$ or $A_j \in \{$"text color is mostly black", "text color is mostly white", "text color is mostly blue", etc.$\}$? Please explain.

4.2 Question 2

Let us examine the assumption that attributes are not anti-correlated with relevance as it relates to the vector space model. Suppose we have the corpus C representing all possible ways that one or more words can be chosen from a vocabulary $V = \{$apple, banana, carrot, date$\}$, as shown below:

- d_1 = "apple"
- d_2 = "banana"
- d_3 = "carrot"
- d_4 = "date"
- d_5 = "apple banana"
- d_6 = "apple carrot"
- d_7 = "apple date"
- d_8 = "banana carrot"
- d_9 = "banana date"
- d_{10} = "carrot date"
- d_{11} = "apple banana carrot"
- d_{12} = "apple banana date"
- d_{13} = "apple carrot date"
- d_{14} = "banana carrot date"
- d_{15} = "apple banana carrot date"

Assume that a document is relevant to a query if it contains each of the words listed in the query.

A query q = "apple" would satisfy the assumption that all terms not contained in a query have the same distribution over relevant documents as they do over non-relevant documents. In fact, it can be shown for this corpus that no matter how long a query is, and no matter which terms it includes, the above assumption regarding correlation with relevance will always be satisfied.

4.2.1 Question 2 Part A

Why is the assumption always satisfied in the above corpus?

4.2.2 Question 2 Part B

What characteristics of the above example make it inapplicable to most real-world applications?

4.2.3 Question 2 Part C

Why did Robertson & Spärck Jones think they were justified in making the assumption that the terms not contained in the query have the same distribution over relevant documents as they do over non-relevant documents in their derivation of a ranking function? Please try to keep your answer as general as possible so that it applies to the probabilistic retrieval model and not simply the vector space model.

4.3 Question 3

Why do you think Robertson & Spärck Jones decided to base the uncertainty in $\Pr(R_q = y \mid \vec{A} = \vec{d})$ for a particular document d on the notion that the set of attributes A is not uniquely specifying? Would it have been just as easy for them to base this uncertainty on the variation among users or within particular users? Why or why not?

5 Sample Answers

5.1 Answer 1

5.1.1 Answer 1 Part A

The assumption was necessary so that we could decompose

$$\prod_{j=1}^{m} \frac{\Pr(A_j = d[j] \mid R_q = y)}{\Pr(A_j = d[j])}$$

into

$$\prod_{j:q[j]\neq 0} \frac{\Pr(A_j = d[j] \mid R_q = y)}{\Pr(A_j = d[j])} \times \prod_{j:q[j]=0} \frac{\Pr(A_j = d[j] \mid R_q = y)}{\Pr(A_j = d[j])}. \tag{7}$$

Recall that being able to decompose the formula in this manner allowed us to think about attributes in terms of whether they are represented in a particular query. We could eliminate the entire second product(relating to non-query attributes) by making the assumption that attributes that do not appear in the query have the same distribution over all documents as they do over relevant documents. This assumption will be examined further in Question 2.

5.1.2 Answer 1 Part B

Requiring a "not on" attribute value for each attribute does not prevent us from representing each of the above attributes as a collection of conditionally dependent attributes. For example, one might think that although an attribute $A_1 = \{$"document is handwritten", "documents is typed"$\}$ would not apply to our model directly, we could break A_1 up into two attributes A_2, and A_3, where $A_2 = \{$"document is handwritten", "document is not handwritten"$\}$, and $A_3 = \{$"document is typed", "document is not typed"$\}$. Clearly, both

A_2 and A_3 have an off value. The problem we have now is that attributes A_2 and A_3 are conditionally dependent on each other (i.e. if A_2 is "on", then A_3 must be "off"). Of course, this constitutes a further violation of any assumptions of independence between attributes.

5.2 Answer 2

5.2.1 Answer 2 Part A

The assumption is so easily satisfied for the example corpus because the probability of a term being in a document is completely independent of the probability of any other term (or set of terms) being in the document.

5.2.2 Answer 2 Part B

This independence between terms (or more generally, attributes) does not apply to a real world corpus. For example, two terms with a similar meaning tend to have a better chance of appearing in the same document as two randomly chosen terms. Without this independence, analyzing whether it is realistic to assume attributes are anti-correlated with relevance becomes a much less trivial task. It is also important to note that most real world applications are based on a more complex notion of relevance than we have used in this example.

5.2.3 Answer 2 Part C

We were able to still make the above assumption because we assumed that a query was "limited" relative to the size of the corpus and attribute set. If a query was not overly expressive, one could argue that attributes not included in the query are just about as likely to occur in any document as they are to occur in a relevant document. As we make our query more expressive, this assumption becomes less appropriate.

5.3 Answer 3

Robertson & Spärck Jones decided to base the uncertainty in a particular document on the notion that the set of attributes A is not uniquely specifying because it is easier to gather statistics for and model documents than it is to gather statistics for and model users. A document is much easier to characterize than the preference of users (or the changing preference of a single user).

References

[1] Karen Spärck Jones and Steve Walker and Stephen Robertson. *A probabilistic model of information retrieval: Development and comparative experiments.* 2000: Information Processing and Management (779-808, 809-840).

[2] William S. Cooper. *Some inconsistencies and misidentified modeling assumptions in probabilistic information retrieval.* 1995: ACM Transactions on Information Systems (TOIS) (100-111).

[3] Michael Gordon and Peter Lenk. *When is the probability ranking principle suboptimal?* 1992: Journal of the American Society for Information Science, 43 (1-14).

[4] Stephen Robertson and Karen Spärck Jones. *Relevance weighting of search terms.* 1976: Journal of the American Society for Information Science, 27 (129-146).